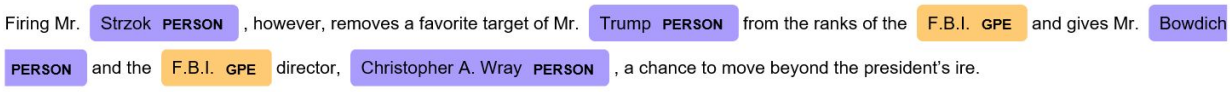
Medical Entity Recognition

Team No 5
Geethika Dudyala(20161154)
Aditya Saripalli(20173071)
Aman Bansal(20161008)
Tarun Vatwani(2018201075)

Project Statement

Medical Entity Recognition(MER) is a subset of a very famous problem in the field of Information Extraction(IE) i.e. Named Entity Recognition(NER). In other words it is a domain specific NER in this case the domain is medical field.
NER is a method/model which takes a series of text(sentences or paragraphs) and provides labels to noun phrases present in that piece of text. The image below shows an example of how NER model typically works.



This project aims at parsing named entities and in this project, we have to recognize and classify medical data into the relevant categories, namely drugs, diseases, symptoms, side-effects, treatment, etc. Twitter data will be the input and based on previous medical data from databases and ontologies, relevant medical terms have to be parsed and classified (medical named entities are recognized and classified based on the category they belong to (ex: drug or a disease or cure etc....)
As the name suggests, a Medical Name Entity Recognizer identifies medical entities in text. Medical entities, in the context of our project, are fixed, there are 5 categories as mentioned above. Previously, researchers in the field have used hand crafted features to identify medical entities in medical literature. In this project, we have to extend medical entity recognition on tweets. We would use NLP toolkits designed for processing tweets along with other medical ontologies (or databases) to exploit semantic features for this task.

Dataset

1. **CADEC dataset :** It is a corpus of medical forum posts on patient-reported Adverse Drug Events (ADEs).
    - ★ The corpus is sourced from posts on social media, and contains text that is largely written in colloquial language and often deviates from formal English grammar and punctuation rules.
    - ★ Annotations contain mentions of concepts such as drugs, adverse effects, symptoms, and diseases linked to their corresponding concepts in controlled vocabularies.
2. **Micromed Dataset :** This dataset was described in MedInfo 2015 paper from IBM Melbourne Research lab.
    - ★ consists of tweet annotations with medical entities. (three types of entities: Disease (T047 in UMLS), Symptoms (T184), and Pharmacologic Substance (T121)

The above two datasets are already available but these datasets are very small in size, hence may not be sufficient for training.
We have to increase the size of the dataset that is available for training heuristically.

For this, we would be given three resources,

R1 : A list of hashtags, which are relevant to the medical domain
R2 : A general tweet corpus (about 40-50 GB in size)
R3 : A list of medical terms and their appropriate categories.

Using the list of hashtags (R1) we can obtain a new dataset consisting of a subset of the general tweet corpus (R2) - medical domain related tweets. This new dataset can be used both for training as well as testing purpose.
A part of this new dataset can be annotated using R3 and could be used for training purposes. The rest of the dataset can be used for testing purposes.

Applications

1. From the results obtained, we can get the specific details of any disease that has widely spread in a particular area.
2. Results could be analysed to find the the patient's feedback/response for a particular drug, the effectiveness of a particular drug (how far it has been successful in treatment and what are the negative points)
3. Results can be utilised by companies producing medical products for improving their sales.

Challenges

1. As the rules and features for the medical data would be different from that of ordinary data, this problem becomes more challenging when compared to named-entity-recognition problem on normal data.
2. Twitter data is user-generated social media text, thus, it would be highly disorganized and prone to inconsistencies. They contain a lot of noise apart from the required information. Filtering the noise/inconsistencies out from the tweets is a major challenge for our project as they could affect the performance drastically.
3. Learning distributed representations for medical tweets. (this can overcome the weaknesses of 'bag-of-words' models)
4. We would have to identify the relevant content from a given tweet. (For example, all tweets containing the keyword 'morphine' might not about the drug 'morphine')
5. Entity linking for exploiting semantic features from ontologies

Implementation Details

MER can be implemented as a sequence classification task, where every chunk is predicted IOB-style as Drug, Disease, Symptom, Treatment and Test. The IOB format (short for inside, outside, beginning) is a common tagging format for tagging tokens. The B- prefix before a tag indicates that the tag is the beginning of a chunk, and an I- prefix before a tag indicates that the tag is inside a chunk. The B- tag is used only when a tag is followed by a tag of the same type without O tokens between them. An O tag indicates that a token belongs to no chunk.

The following example indicates the IOB-style tagging:
Sentence: *"Insulin is prescribed for the type-2 diabetes"*
IOB tags: *"Insulin: B-drug, is: O, prescribed: O, for: O, the: O, type-2: B-disease, diabetes: I-disease"*

Milestones

1. Studying and Implementing basic sequence to sequence models such as LSTM for NER
2. Study state of the art methods sequential models and try to test them for twitter data set
3. Working on twitter corpus to create a training set for tweets using R1,R2 and R3
4. Final documentation and PPT
5. Preparation of working demo

Tools:

Toolkits that we would be using :

1. ***CRF++*** : Our named entity model would be modeled with a CRF model.
2. ***NLTK*** : It is a suite of libraries for NLP, in python.
3. ***Metamap*** : To map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques.

References:

1. https://www.ncbi.nlm.nih.gov/pubmed/25817970
2. https://metamap.nlm.nih.gov/
3. https://github.com/IBMMRL/medinfo2015
4. CliNER: https://arxiv.org/abs/1803.02245
5. SciBERT: https://arxiv.org/pdf/1903.10676v3.pdf
6. BioFLAIR: https://arxiv.org/pdf/1908.05760v1.pdf
7. CollaboNET: https://arxiv.org/pdf/1809.07950v2.pdf
8. Clinical Concept Extraction: https://arxiv.org/abs/1810.10566